FINAL EXAM - CSC 262 - May 6, 2015

NAME: ____

ID: ____

Answer all 10 questions in the space provided. All questions have equal weight. Use the back of the page for additional space if needed (please indicate clearly where you have done this). Critical values for the standard normal, t, χ^2 and F distributions will be given as needed. Probabilities associated with any other distributions may be evaluated using normal approximation methods. A methodological summary is included with the exam. A calculator is permitted. The exam will last 3 hours.

Q1: For an *iid* sample of size $n = 26$ from a normal distribution $N(\mu, \sigma^2)$ we are given sample mean $\bar{X} = 11.56$ and sample standard deviation $S = 2.32$.

(a) Is there sufficient evidence to reject the null hypothesis $H_o : \mu = 12$ in favor of the two-sided alternative hypothesis $H_a : \mu \neq 12$ with a significance level of $\alpha = 0.05$?

(b) Calculate a confidence level $1 - \alpha = 0.95$ upper bound for σ.

(c) Using the upper bound for σ calculated in part (b) estimate the sample size needed to obtain a confidence interval for μ with a margin of error of 0.5. Assume the sample size will large. Use confidence level $1 - \alpha = 0.95$.

[$t_{25,0.025} = 2.06$, $\chi^2_{25,0.95} = 14.61$, $z_{0.025} = 1.96$]

Q1: [Answer]

(a)

$$\begin{aligned} T &= \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \\ &= \frac{11.56 - 12}{2.32/\sqrt{26}} \\ &= -0.967. \end{aligned}$$

Note that $|T| < t_{25,0.025} = 2.06$, so do not reject H_o at $\alpha = 0.05$.

(b)

$$\begin{aligned} UB &= \frac{S}{\sqrt{\chi^2_{n-1,1-\alpha}/(n-1)}} \\ &= \frac{2.32}{\sqrt{14.61/25}} \\ &= 3.03. \end{aligned}$$

So,

$$\sigma < 3.03$$

is the 95% upper confidence bound for σ.

(c) Use estimate $\hat{\sigma} = 3.03$ in formula

$$n \approx \left(z_{\alpha/2} \frac{\hat{\sigma}}{E_o} \right)^2 = \left(1.96 \times \frac{3.03}{0.5} \right)^2 = 141.51,$$

so round up to $n = 142$.

Q2: We are given two independent samples from normally distributed populations. The data is summarized in the table below.

(a) Use an F-test to test for equality of variances, using significance level $\alpha = 0.05$.

(b) Using the appropriate procedure based on the test for equality of variances (that is, either a pooled variance t-test or Welch's t-test), calculate a confidence interval for $\mu_1 - \mu_2$ with confidence level $1 - \alpha = 0.95$.

$[F_{4,9,0.975} = 0.112,\ F_{4,9,0.025} = 4.718,\ t_{11,0.025} = 2.201]$

	Sample 1	Sample 2
$\bar{X}_i$	12.2780	17.5310
S_i	0.1730	0.6820
n_i	5	10

Q2: [Answer]

(a) We have

$$F = \frac{0.1730^2}{0.6820^2} = 0.0643.$$

Reject $H_o : \sigma_1^2 = \sigma_2^2$ if F is less than or equal to $F_{1-\alpha/2,n_1-1,n_2-1} = 0.112$ or if F is greater than or equal to $F_{\alpha/2,n_1-1,n_2-1} = 4.718$, where $\alpha = 0.05$. Therefore, reject the null hypothesis of equal variances at a significance level $\alpha = 0.05$.

(b) We conclude that $\sigma_1^2 \neq \sigma_2^2$, so use Welch's procedure for unequal variances. The degrees of freedom is given by

$$\begin{aligned}
\nu_W &= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}} \\
&= \frac{\left(\frac{0.1730^2}{5} + \frac{0.6820^2}{10}\right)^2}{\frac{(0.1730^2/5)^2}{5-1} + \frac{(0.6820^2/10)^2}{10-1}} \\
&= 11.05362.
\end{aligned}$$

Round down to $\nu_W = 11$ degrees of freedom. The confidence interval is then

$$\begin{aligned}
CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{\nu_W,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\
&= \bar{X}_1 - \bar{X}_2 \pm t_{11,0.025}\sqrt{\frac{0.1730^2}{5} + \frac{0.6820^2}{10}} \\
&= 12.2780 - 17.5310 \pm 2.201 \times 0.229 \\
&= -5.253 \pm 0.504 \\
&= (-5.757, -4.749).
\end{aligned}$$

The confidence interval is (-5.76,-4.75) or -5.25 ± 0.504.

Q3: We are given two paired samples from normally distributed populations ($n = 5$). The data is summarized in the table below. Perform a two-sided hypothesis test using hypotheses $H_o : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$. Use significance level $\alpha = 0.05$.

	Sample 1	Sample 2	Difference
1	139.6	146.4	-6.8
2	139.3	140.7	-1.4
3	141.4	142.6	-1.2
4	139.1	145.6	-6.5
5	136.1	145.4	-9.3

[$t_{4,0.025} = 2.776$]

Q3: [Answer] We have

$$\begin{aligned} \bar{X}_1 &= 139.1 \\ \bar{X}_2 &= 144.14 \\ \bar{X}_1 - \bar{X}_2 &= -5.04 \\ S_D &= 3.584. \end{aligned}$$

Test statistic is

$$\begin{aligned} T &= \frac{\bar{D}}{S_D/\sqrt{n}} \\ &= \frac{\bar{X}_1 - \bar{X}_2}{S_D/\sqrt{n}} \\ &= \frac{-5.04}{3.584/\sqrt{5}} \\ &= -3.14. \end{aligned}$$

Reject H_o if

$$|T| \geq t_{n-1,\alpha/2} = t_{4,0.025} = 2.776.$$

Therefore, reject the null hypothesis at a significance level $\alpha = 0.05$.

Q4: A coin tossing procedure is tested for fairness. Out of $n = 25$ tosses $X = 11$ result in *heads*.

(a) Let p be the true probability of tossing *heads*. Test the null hypothesis $H_o : p = 0.5$ against the two-sided alternative $H_a : p \neq 0.5$. Use significance level $\alpha = 0.01$. Use a normal approximation with continuity correction.

(b) Suppose we wish to construct a confidence interval for p with confidence level $1 - \alpha = 0.99$ and a margin of error no greater than $E = 0.025$. What sample size would be needed?

[$z_{0.005} = 2.576$]

Q4: [Answer]

(a) Under the null hypothesis, $X \sim bin(25, 1/2)$, with mean and variance $\mu = np = 25 \times 1/2 = 12.5$ and $\sigma^2 = np(1-p) = 25 \times 1/2 \times 1/2 = 6.25$. Suppose $X_{norm} \sim N(\mu = 12.5, \sigma^2 = 6.25)$. Then

$$\alpha_{obs} = 2P(X \leq 11) \approx 2P(X_{norm} \leq 11 + 0.5) = 2P(X_{norm} \leq 11.5).$$

Equivalently, we have z-score

$$Z = \frac{11.5 - \mu}{\sigma} = \frac{11.5 - 12.5}{\sqrt{6.25}} = -0.4.$$

Reject H_o if $|Z|$ is greater than or equal to $z_{\alpha/2} = 2.576$. Therefore, do not reject the null hypothesis at a significance level $\alpha = 0.01$.

(b) If $E = 0.025$, use as estimate $p^* = 0.5$. We have $z_{\alpha/2} = 2.576$, so

$$n = p^*(1-p^*)\left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.25\left(\frac{2.576}{0.025}\right)^2 = 2653.959.$$

Round up to $n = 2654$.

Q5: A certain experimental cancer therapy was evaluated in a clinical trial. A control group of 70 subjects was given standard care, while a treatment group of 50 subjects was given the experimental therapy. The subjects were observed for a 5 year period. Of the control group, 29 subjects experienced recurrence, while of the treatment group, 15 subjects experienced recurrence.

	Control ($i = 1$)	Treatment ($i = 2$)
X_i = Number of recurrences	29	15
n_i = Group sample size	70	50
$\hat{p}_i$ = Observed recurrence rate	0.414	0.3

(a) Construct a confidence interval for the difference in recurrence rates $p_2 - p_1$ between the Treatment and Control groups. Use confidence level $1 - \alpha = 0.95$.

(b) Construct a confidence interval for the log odds ratio:

$$\log \left[\frac{Odds(Recurrence \mid Treatment)}{Odds(Recurrence \mid Control)} \right].$$

Use confidence level $1 - \alpha = 0.95$.

(c) What can be concluded from parts (a) and (b) as to whether or not recurrence rates differ between the groups? Are the conclusions from parts (a) and (b) consistent?

$[z_{0.025} = 1.96]$

Q5: [Answer]

(a) We have estimates $\hat{p}_1 = 0.248$, $\hat{p}_2 = 0.306$. The confidence interval is then

$$\begin{aligned} CI_{1-\alpha} &= \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= 0.3 - 0.414 \pm 1.96\sqrt{\frac{0.3(1-0,3)}{70} + \frac{0.414(1-0.414)}{50}} \\ &= -0.114 \pm 0.172 \\ &= (-0.286, 0.0573). \end{aligned}$$

(b) We have contingency table

	Control	Treatment
Recurrence	29	15
No Recurrence	41	35

The estimate of the OR is

$$\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 29 \times 35/(41 \times 15) = 1.65.$$

The standard error is

$$SE(\log(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{29} + \frac{1}{15} + \frac{1}{41} + \frac{1}{35}} = 0.3926.$$

An approximate $(1-\alpha)100\%$ confidence interval for $\log(OR)$ is therefore

$$\begin{aligned} CI_{1-\alpha} &= \log(\hat{OR}) \pm z_{\alpha/2} SE(\log(OR)) \\ &= \log(1.65) \pm 1.96 \times 0.3926 \\ &= 0.501 \pm 0.769 \\ &= (-0.268, 1.270). \end{aligned}$$

(c) From part (a), the CI for $p_2 - p_1$ contains 0. From part (b) the CI for log OR also contains 0, equivalent to OR = 1. Both conclusions are consistent with the null hypothesis $H_o : p_1 = p_2$ (and so are consistent with each other).

Q6: In statistical genetics a *dihybrid cross* produces phenotype pairs in the ratio 9:3:3:1. Suppose for this type of cross in a certain plant, the phenotypes Tall/Yellow, Short/Yellow, Tall/Green, Short/Green are expected to conform to these ratios. Suppose a sample of $n = 120$ of this cross yields the following counts:

	Tall/Yellow	Short/Yellow	Tall/Green	Short/Green	Totals
Observed counts O_i	68	22	20	10	120
Hypothetical frequencies p_i	9/16	3/16	3/16	1/16	1
Expected counts					120

(a) For each phenotype pair, calculate the expected count assuming that the hypothetical frequencies are true. Place you answers in the table above.

(b) Perform a χ^2 test against the null hypothesis $H_o : p_i$ are the true population frequencies. Use significance level $\alpha = 0.05$. Yate's correction is not needed.

[$\chi^2_{3,0.05} = 7.815$]

Q6: [Answer]

(a) The expected counts $E_i = np_i$ are given in the following table:

	Tall/Yellow	Short/Yellow	Tall/Green	Short/Green	Totals
Observed counts O_i	68	22	20	10	120
Expected counts E_i	67.5	22.5	22.5	7.5	120.0
$(O_i - E_i)^2/E_i$	0.0037	0.0111	0.2778	0.8333	1.125926

(b) Without Yate's correction

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \approx 0.0037 + 0.0111 + 0.2778 + 0.8333 = 1.126.$$

Reject H_o if X^2 is greater than or equal to $\chi^2_{k-1,\alpha} = 7.815$. Therefore, do not reject the null hypothesis at a signficance level $\alpha = 0.05$.

Q7: A contingency table with $n_r = 3$ rows and $n_c = 3$ columns based on a random sample of size $n = 200$ is given below. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Perform a χ^2 test against the null hypothesis of row and column independence $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$.

Observed counts $O_{i,j}$

	1	2	3	Totals
1	20	21	51	92
2	12	21	6	39
3	29	37	3	69
Totals	61	79	60	200

$[\chi^2_{4,0.05} = 9.488]$

Q7: [Answer] The expected counts $E_{i,j} = R_i C_j / n$ are given in the following table, where R_i, C_j are the row and column totals. For example,

$$E_{1,3} = \frac{R_1 C_3}{n} = \frac{92 \times 60}{200} = 27.6.$$

Expected counts $E_{i,j}$

	1	2	3	Totals
1	28.06	36.34	27.60	92.00
2	11.89	15.40	11.70	39.00
3	21.05	27.25	20.70	69.00
Totals	61.00	79.00	60.00	200.00

X^2 statistic terms $(O_{i,j} - E_{i,j})^2 / E_{i,j}$

	1	2	3	Totals
1	2.32	6.48	19.84	28.63
2	0.00	2.03	2.78	4.81
3	3.01	3.48	15.13	21.63
Totals	5.32	11.99	37.75	55.07

Reject H_o if X^2 is greater than or equal to $\chi^2_{(n_r-1)(n_c-1),\alpha} = \chi^2_{4,0.05} = 9.488$. Without Yate's correction, we have

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \approx 55.066.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$. Note that is suffices to calculate only

$$\frac{(O_{1,3} - E_{1,3})^2}{E_{1,3}} = 19.84 \text{ or } \frac{(O_{3,3} - E_{3,3})^2}{E_{3,3}} = 15.13,$$

since we could then conclude that $X^2 \geq \chi^2_{4,0.05}$, without the need to calculate X^2 (since all other terms are positive).

Q8: We are given two paired samples of sample size $n = 6$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a lower tailed sign test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D < 0$. Use significance level $\alpha = 0.05$."

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	10.5	15.7	-5.2	
2	8.2	11.5	-3.3	
3	12.0	12.4	-0.4	
4	14.5	11.6	2.9	
5	8.3	12.0	-3.7	
6	4.0	4.0	0.0	

[$z_{0.05} = 1.645$]

Q8: [Answer] After excluding ties there are $X = 1$ positive differences among $n' = 5$ pairs. Using the **binomial distribution** directly:

$$\alpha_{obs} = P(X \leq 1) = P(X = 0) + P(X = 1) = 0.5^5 + 5 \times 0.5^5 = 0.1875.$$

So, do not reject H_o with $\alpha = 0.05$ significance level.

Using a **normal approximation**, under the null hypothesis, $X \sim bin(5, 1/2)$, with mean and variance $\mu = np = 5 \times 1/2 = 2.5$ and $\sigma^2 = np(1-p) = 5 \times 1/2 \times 1/2 = 1.25$. Then we have z-score

$$\begin{aligned} Z &= \frac{1-\mu}{\sigma} = \frac{1-2.5}{\sqrt{1.25}} = -1.34 \text{ without continuity correction,} \\ Z &= \frac{1.5-\mu}{\sigma} = \frac{1.5-2.5}{\sqrt{1.25}} = -0.89 \text{ with continuity correction.} \end{aligned}$$

In either case, we reject H_o if $Z \leq -z_\alpha = -1.645$. So, do not reject H_o with $\alpha = 0.05$ significance level.

Q9: We are given two paired samples of sample size $n = 8$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a two-sided signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	97.9	94.7	3.2	
2	104.7	107.0	-2.3	
3	91.9	90.3	1.6	
4	96.7	94.1	2.6	
5	106.4	103.3	3.1	
6	103.4	100.1	3.3	
7	94.2	89.9	4.3	
8	97.7	102.1	-4.4	

[$z_{0.025} = 1.96$]

Q9: [Answer] The signed ranks are given in the following table:

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\lvert D\rvert$	Sign
1	97.9	94.7	3.2	5.0	+
2	104.7	107.0	-2.3	2.0	-
3	91.9	90.3	1.6	1.0	+
4	96.7	94.1	2.6	3.0	+
5	106.4	103.3	3.1	4.0	+
6	103.4	100.1	3.3	6.0	+
7	94.2	89.9	4.3	7.0	+
8	97.7	102.1	-4.4	8.0	-

There are no ties, so $n = 8$ pairs. The negative and positive rank sums are, respectively,

$$T_- = 10 \text{ and } T_+ = 26.$$

Then

$$T_{obs} = \min(T_-, T_+) = \min(10, 26) = 10.$$

The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = \frac{n(n+1)}{4} = 18 \text{ and } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = 7.141.$$

This gives z-score

$$\begin{aligned} Z &= \frac{T_{obs} - \mu_T}{\sigma_T} = \frac{10 - 18}{7.141} \approx -1.12 \text{ without continuity correction,} \\ Z &= \frac{T_{obs} + 0.5 - \mu_T}{\sigma_T} = \frac{10.5 - 18}{7.141} \approx -1.05 \text{ with continuity correction.} \end{aligned}$$

In either case, we reject H_o if $Z \leq -z_{\alpha/2} = -1.96$. So, do not reject H_o with $\alpha = 0.05$ significance level.

Q10: We are given two independent samples of sample sizes $n_1 = 5$, $n_2 = 10$. The data is summarized in the table below. Suppose $\tilde{\mu_i}$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu_1} - \tilde{\mu_2} = 0$ against $H_a : \tilde{\mu_1} - \tilde{\mu_2} \neq 0$. Use significance level $\alpha = 0.05$.

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	30.0	23.1	25.8	25.9	23.5						25.8
Sample 2	22.3	23.2	20.3	21.7	25.6	22.9	22.9	20.7	22.3	19.0	22.3

$[z_{0.025} = 1.96]$

Q10: [Answer] The signed ranks are given in the following table:

	1	2	3	4	5	6	7	8	9	10	$\tilde{X}_i$
Sample 1	30.0	23.1	25.8	25.9	23.5						25.8
Sample 2	22.3	23.2	20.3	21.7	25.6	22.9	22.9	20.7	22.3	19.0	22.3
Ranks 1	15.0	9.0	13.0	14.0	11.0						0.0
Ranks 2	5.5	10.0	2.0	4.0	12.0	7.5	7.5	3.0	5.5	1.0	0.0

The rank sums for samples 1 and 2 are, respectively,

$$T_1 = 62 \text{ and } T_2 = 58.$$

We only need T_1. The mean and standard deviation of T_1 is

$$\mu_{T_1} = \frac{n_1(n_1+n_2+1)}{2} = 40 \text{ and } \sigma_{T_1} = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}} = 8.165.$$

This gives z-score (noting that $T_1 > \mu_{T_1}$),

$$\begin{aligned} Z &= \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{62-40}{8.165} \approx 2.694 \text{ without continuity correction,} \\ Z &= \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{62-0.5-40}{8.165} \approx 2.633 \text{ with continuity correction.} \end{aligned}$$

In either case, we reject H_o if $Z \geq z_{\alpha/2} = 1.96$. So, reject H_o with $\alpha = 0.05$ significance level.